DEEP BLIND SYNTHESIZED IMAGE QUALITY ASSESSMENT WITH CONTEXTUAL MULTI-LEVEL FEATURE POOLING

Xiaochuan Wang[1], *Kai Wang*[1], *Bailin Yang*[2], *Frederick W.B. Li*[3], *and Xiaohui Liang*[1]*

State Key Laboratory of VRTS, Beihang University[1]
School of Computer Science and Information Engineering, Zhejiang Gongshang University[2]
Department of Computer Science, University of Durham[3]

ABSTRACT

Blind image quality metrics have achieved significant improvement on traditional 2D image dataset, yet still being insufficient for evaluating synthesized images generated from depth-image-based rendering. The geometric distortions in synthesized image are non-uniform, which is challenging for feature representation and pooling. To address this, we propose an end-to-end deep blind synthesized image quality metric SIQA-CFP. We particularly design a contextual multi-level feature pooling module to encode low- and high-level features, which are extracted by a deep pre-trained ResNet. Experimental results on IRCCyN/IVC DIBR dataset show that our method outperforms state-of-the-art synthesized image quality metrics. Our method also achieves competitive performance on traditional 2D image datasets like LIVE Challenge and TID2013.

Index Terms— image quality assessment, synthesized image, feature pooling, DIBR, deep learning

1. INTRODUCTION

Depth-image-based rendering (DIBR) has been widely used in 3D applications, such as 3DTV [1] and free-viewpoint video (FVV) [2]. By utilizing a few reference views, it can synthesize arbitrary new virtual views without knowing the ground truth. However, such synthesized images suffer from distortions, especially the geometric distortion, leading to inferior quality of experience (QoE). To maintain the quality of service (QoS), being able to evaluate synthesized image quality thus becomes an emergency.

Traditional image distortions, such as Gaussian blurring, white noise and blocking artifact, distribute homogeneously across a distorted image. In contrast, as depth discontinuities occurring around the disoccluded regions of a synthesized view, it leads to locally and non-uniformly distributed geometric distortions as illustrated in Fig. 1. The sensitivity of such local structure distortion is related to its contextual regions, i.e., luminance adaptation or contrast masking of its neighborhoods, as validated by psychology and cognitive sciences [3]. Feature representation and pooling of synthesized images therefore become challenging, where previous 2D image quality assessment (IQA) methods designed for traditional distortions are incapable to deal with such distortion.

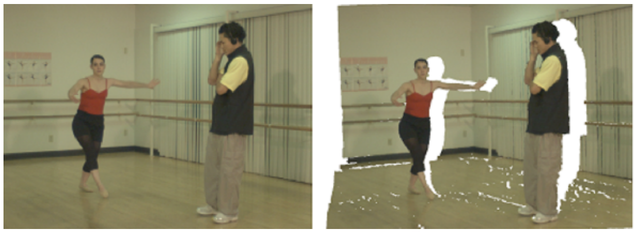


Fig. 1. Reference image and a DIBR synthesized image. The blank regions indicate the geometric distortions.

Recently, deep learning has attracted a great attention in computer vision tasks, where the convolutional neural network (CNN) is utilized to represent image features rather than relying on handcraft ones. Despite of its success in general computer vision tasks, applying deep learning to blind synthesized image quality assessment still encounters difficulties [4]. First of all, current DIBR synthesized image benchmarks are too small to train a deep model against overfitting. Transfer learning with pre-trained model may partially solve the problem. However, a proper feature presentation and pooling strategy are desired for evaluating synthesized image quality.

This paper proposes a novel end-to-end blind synthesized image quality with contextual multi-level feature representation and pooling (SIQA-CMP). We investigate and select low- and high-level outputs from a pre-trained ResNet model on ImageNet classification to represent synthesized image features. We then design a contextual pooling to aggregate the multi-level features, and finally regress to subjective scores. The proposed contextual multi-level feature pooling is believed to properly represent geometric distortions toward the image quality, which is validated by testing against the IRCCyN/IVC [5] and our DIBR image datasets [6]. Notably, our metric can be generalized to traditional IQA databases.

*{liang_xiaohui@buaa.edu.cn} is the corresponding author. Supported by National Nature Science Foundation of China (No. 61572058).

2. RELATED WORK

IQA methods can be divided into full-reference (FR-), reduced-reference (RR-) and no-reference (NR-) according to the knowledge of pristine image [7]. Since the ground truth of synthesized images are usually not available in DIBR-related applications, we only concern about NR-IQA or blind IQA.

Most blind IQA methods predict distorted image by using natural scene statistics (NSS), where natural images are assumed to have a similar statistical distribution in spatial or transform domain. A common framework comprises two stages. A distorted image is firstly represented with NSS features, and then pooled to get the predicted scores. Typical metrics include BRISQUE [8], NIQE [9], etc.

Deep learning has recently been introduced into IQA. Due to the feature extraction power of non-linear convolution, CNN-based NR-IQA methods have outperformed traditional NR-IQAs, approaching the state-of-the-art FR-IQAs. Typical examples include [10, 11, 12, 13], where the representation and pooling of features extracted from CNN model are diverse. Kang [10] proposed using the final output and max/min-pooling layers to train the quality prediction model. Bosse [11] adopted a similar strategy, but improving the assignment of image patch scores with saliency. Kim [12] proposed extracting patch-aware features through convolutional network and spatially pooled them to the subjective scores. More recently, MFIQA [13] proposed using multi-level features extracted from pre-trained deep model, and then pooled them with simple global average pooling.

Current blind synthesized image quality metrics still follow the traditional way, but paying extra attention on geometric distortions. For instance, Tian [14] assumed that holes regions appear differently before and after morphological operations. The distortion degrees can thus be estimated by extracting and pooling the differential map features. Gu [15] proposed using auto-regression model to evaluate the stationarity of local intensity, where the local region containing geometric distortions is assumed to result in high variance. A saliency thresholding is utilized to pool the local features. No deep learning based methods has been reported to deal with DIBR synthesized images.

3. FRAMEWORK

The framework of our blind synthesized image quality assessment with contextual multi-level feature representation and pooling (SIQA-CMP) is illustrated in Fig. 2. Inspired by transfer learning, we exploit the ResNet-50 model pretraining on ImageNet database for image classification to extract multi-level features of synthesized images. We then design a contextual pooling to encode low- and high-level features, with the aim of balancing the contextual effect of local features. Finally, we train the contextual pooling module with our new DIBR synthesized image dataset [6].

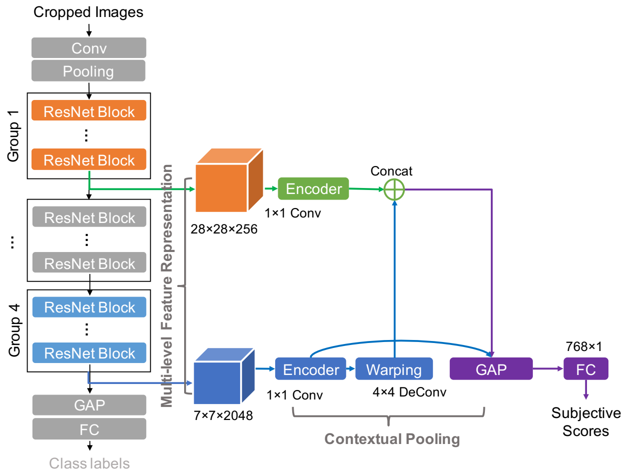


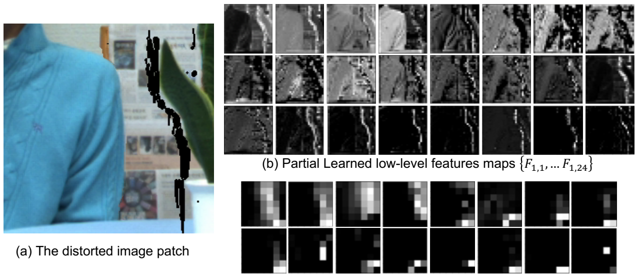
Fig. 2. Overall framework of our proposed metric. ’Conv’, ’DeConv’, ’FC’, ’Concat’ and ’GAP’ indicate convolutional, deconvolutional, fully-connected, concatenate, and global average pooling layers, respectively.

Multi-level feature representation. As investigated by previous work, a deep CNN model is a multi-level feature extractor. Low-level features are learned in the early layers, while in the deeper layers, more abstract information is perceived [16]. We choose the ResNet-50 pre-trained on ImageNet dataset to extract multi-level features [17]. We select the two outputs after the $1st$ and $4th$ residual groups from the ResNet-50 model, denoting as $\{F_{1,n}\}$ and $\{F_{4,m}\}$, where $n \in [1, 256]$, and $m \in [1, 2048]$. The original global average pooling and fully-connected layers after the 4th residual groups are removed.

As the original ResNet model is trained with 224×224 images, we crop each input distorted image into 50 image patches (224×224). We particularly use slide-window strategy to maximize the exploitation of a whole image. The extracted patch-aware features are then pooled through our proposed contextual feature pooling module to predict the quality scores.

Contextual feature pooling. From computational cognitive science, an early convolutional layer has local receptive field with respect to local regions in the image. The receptive field is enlarged in a later convolutional layer, becoming sensitive to larger scale information. The benefit of multi-level feature pooling has been validated in [13], where the multi-level features are separately averaged along the depth dimension, and then simply pooled for regression.

In synthesized images, geometric distortions are non-uniform and related to its surrounding regions. For instance, a small hole yet having structure features would be insensitive if it is masked by complex textures in the surroundings. Examples of extracted features are shown in Fig. 3. We can see



(a) The distorted image patch

(b) Partial Learned low-level features maps $\{F_{1,1}, ... F_{1,24}\}$

(c) Partial learned high-level features maps $\{F_{4,1}, ... F_{4,16}\}$

Fig. 3. Feature maps extracted by using pre-trained ResNet-50 model. (a) is a distorted image patch, while its low- and high-level feature maps are sorted from top to bottom showing on its right. Note that (b) depicts structure features of local regions, while (c) emphasizing the context of holes.

that the structure features of holes are sensitive in low-level features, but being obsolete in the high-level features. An early averaging and a simple pooling cannot represent such contextual relationship properly.

Instead, we firstly aggregate low- and high-level features by warping the high-level features to align the size of the low-level features. Particularly, since $F_{1,n}$ is $28 \times 28 \times 256$ while $F_{4,m}$ is $7 \times 7 \times 2048$, we encode the outputs with a 1×1 convolutional layer, so as to align the channels. We then warp $F_{4,m}$ to the scale of $F_{1,n}$ via a deconvolutional layer. The warped high-level features and the low-level features are aggregated, combining with the pristine $7 \times 7 \times 2048$ high-level features by using the global average pooling [1].

Our contextual pooling is different from previous pooling strategies from two aspects. First, high-level features are aggregated to low-level features with warping. Compared with [13], the contextual information is preserved before the global average pooling. Secondly, we additionally encode the high-level features into the concatenate layer. The main idea is to weight the local structure information and contextual information. It fits for the observation that geometric distortions are contextual-aware.

Database. Considering that current benchmark IRCCyN/IVC DIBR image dataset contains only 3 reference scenes and 84 distorted images [5], we build a new DIBR image dataset, which contains 12 total different reference scenes [6]. Similar to IRCCyN/IVC DIBR image dataset, we warp the reference images to 4 different virtual viewpoints, and then apply 7 DIBR algorithms to generate a total of 336 synthesized images. To validate the performance of the proposed metric, we also test it on IRCCyN/IVC DIBR image dataset, and most popular 2D image quality datasets, LIVE [18] and TID2013 [19]. The LIVE IQA dataset contains only homogeneous distortions, e.g., Gaussian noise, Gaussian blur. TID2013 dataset contains 24 kinds of distortion, where the $14th$ distortion *Non eccentricity pattern noise* and the $15th$ distortion *Local block-wise distortions of different intensity* are similar to geometric distortion, which distribute locally in the images.

Training and testing. To train the metric, we randomly divided our DIBR image dataset into two subsets, training (80%) and testing (20%). The subsets were divided with respect to the reference scenes, so as to prevent overfitting. The mean squared error (MSE) between the predicted image scores and the subjective scores was used as the loss function, with a stochastic gradient optimizer with momentum of 0.9 and initial learning rate of 10^{-3}. The training iterated for 7 epochs, where an early stopping is used to prevent overfitting. During the testing stage, the distorted image was firstly cropped into patches. The predicted patch scores were averaged to produce the final quality score. To testify the proposed metric, we used three standard evaluation indices, i.e., Spearman's rank order correlation coefficient (SROCC), Pearson's linear correlation coefficient (PLCC) and Root mean squared error (RMSE).

4. ABLATION STUDY

To analyze the effectiveness of proposed feature representation and contextual pooling strategy, we tested them independently. Firstly, we substituted the pre-trained ResNet-50 with another pre-trained model VGG-19 on ImageNet dataset [20], while the outputs of *Conv3_4* and *Conv5_4* are exploited as low- and high-level features, respectively (Row 1 of Table 1). We then tested each feature level separately, where the extracted features were directly averaged and pooled (Row 2 and 3 of Table 1). Thirdly, we tested the performance of inverse feature aggregation by warping the low-level features to the scale of the high-level features (Row 4 of Table 1). Finally, we reused the two level features from the ResNet model, but removing the contextual pooling (Row 5 of Table 1). The extracted multi-level features were directly averaged and pooled. The results are shown in Table 1. Each result is the average of ten-fold-cross-validation with random training and testing subsets. The blue tests indicate the highest performance among each group.

We can draw three conclusions from the table. First, the backbone network benefits the performance, i.e., the ResNet-50 with deeper and advanced network architecture is superior than VGG-19. Secondly, the single-level features are insufficient for evaluating synthesized image quality. The individual low-level features or high-level features with straight pooling is inferior than our method, which validates the reasonability of contextual pooling to some extent. In paticular, the high-level features followed by simple pooling like previous work [10, 11, 12] performs even inferior than that with only low-level features, indicating that contextual information conceals distortions. Finally, the inverse warping is inferior than our

[1] https://github.com/wangxiaochaun/Contextual-Multi-level-Feature-pooling-for-synthesized-image-quality-assessment

Table 1. Performance comparison of different strategies. '✓' indicates the parameters in the CNN model are being updated, while other rows are results without parameter updating.

Update back-bone	Strategy	SROCC	PLCC	RMSE
✓	VGG-19	0.872	0.908	0.099
-	$\{F_{1,n}\}$	0.878	0.910	0.107
-	$\{F_{4,m}\}$	0.826	0.826	0.124
✓	inv. warping	0.865	0.909	0.094
✓	w/o c.f.p	0.878	0.910	0.093
-	Ours	**0.879**	**0.938**	**0.065**

Table 2. SROCC and PLCC comparison on our DIBR dataset and the IRCCyN/IVC DIBR image dataset. Italics indicate CNN-based methods.

Type		Our dataset		IRCCyN/IVC	
		SROCC	PLCC	SROCC	PLCC
NR,2D	BRISQUE	0.312	0.301	0.320	0.307
NR,2D	NIQE	0.109	0.102	0.118	0.115
NR,2D	*Kang*	0.240	0.295	0.281	0.302
NR,2D	*MFIQA*	0.156	0.510	0.346	0.345
FR,DIBR	3DSwiM	0.612	0.632	0.616	0.662
FR,DIBR	SDRD	0.742	0.788	0.810	0.761
NR,DIBR	NIQSV+	0.662	0.711	0.667	0.711
NR,DIBR	APT	0.708	0.725	0.716	0.730
NR,DIBR	*Ours*	**0.879**	**0.938**	**0.967**	**0.976**

method, stating that emphasizing local information is against the perception of geometric distortion. Additionally, we utilize all the features without the proposed contextual pooling, as applied in [13]. The results validate that the early GAP and simple pooling is insufficient for evaluating synthesized image quality.

5. BENCHMARK

We benchmarked four NR-IQA metrics (BRISQUE [8], NIQE [9], Kang [10], and MFIQA [13]) designed for traditional 2D image, and four DIBR-related metrics (3DSwiM [21] and SDRD [22] as FR-IQAs, NIQSV+ [14] and APT [15] as NR-IQAs). For the CNN-based metrics, we adopted the same training and testing on our DIBR image dataset. The trained models are evaluated on the IRCCyN/IVC DIBR image dataset. The results are shown in Table 2. Our metric achieved higher correlation scores than the previous methods. Particularly, the performance outperforms previous FR-IQAs.

Table 3. SROCC comparision on traditional image datasets.

Train	Test	Bosse [11]	MFIQA	Ours
LIVE	LIVE	0.956	0.964	**0.986**
TID2013	TID2013	0.882	0.240	**0.910**

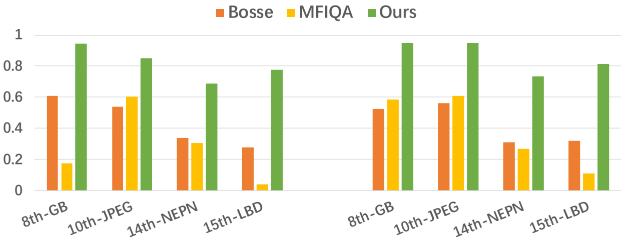


Fig. 4. SROCC and PLCC comparison of individual distortion types on the TID2013 dataset. First four groups are SROCC, next four are PLCC. The partially shown distortion types are *8th-Gaussion blur, 10th-JPEG compression, 14th-Non eccentricity pattern noise and 15th-Local block-wise distortions of different intensity*.

As depicted in Table 3, we additionally trained and tested the proposed metric on LIVE and TID2013 datasets, showing our metric achieved significant improvement comparing to previous CNN-based methods. Particularly, we separately tested the correlation scores on the $14th$ and $15th$ distortions in the TID2013. As in Fig. 4, we can find the different behaviors of feature representation and pooling. Generally, multi-level features representation (MFIQA and Ours) performs better than single outputs and straight pooling ([11]), while the contextual pooling outperforms existing pooling strategies, e.g., max-/min-pooling ([10, 11]) and early GAP [13], especially on those local-distributed distortions.

6. CONCLUSION

We proposed a novel end-to-end blind synthesized image quality metric SIQA-CFP. By analyzing of geometric distortion and investigating effectiveness of features extracted from pre-trained CNN model, we proposed using multi-level feature to represent local and contextual information. We particularly designed a contextual pooling to aggregate the low- and high-level features. The proposed method outperforms previous work both on our DIBR dataset and the benchmark IRCCyN/IVC DIBR image dataset. The contextual multi-level feature pooling can also benefit the performance on traditional image dataset, especially on the locally-distributed distortion types.

7. REFERENCES

[1] Christoph Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5291, pp. 93–104, 2004.

[2] Aljoscha Smolic, "3D video and free viewpoint video from capture to display," *Pattern Recognition*, vol. 44, no. 9, pp. 1958–1968, 2011.

[3] XK Yang, WS Ling, ZK Lu, Ee Ping Ong, and SS Yao, "Just noticeable distortion model and its applications in video coding," *Signal Processing: Image Communication*, vol. 20, no. 7, pp. 662–680, 2005.

[4] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, Nov 2017.

[5] IRCCyN/IVC DIBR Image dataset, "(online): available at http://www.irccyn.ec-nantes.fr/spip.php?article865," June 2009.

[6] VRTS DIBR Image dataset, "(online): available at https://github.com/wangxiaochaun/vrts_dibr_image_dataset," 2018.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] A Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain.," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[9] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[10] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[11] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek, "A deep neural network for image quality assessment," in *IEEE International Conference on Image Processing*, 2016, pp. 3773–3777.

[12] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1969–1977.

[13] Jongyoo Kim, Anh-Duc Nguyen, Sewoong Ahn, Chong Luo, and Sanghoon Lee, "Multiple level feature-based universal blind image quality assessment model," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 291–295.

[14] Shishun Tian, Lu Zhang, Luce Morin, and Olivier Déforges, "Niqsv+: A no-reference synthesized view quality assessment metric," *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[15] K. Gu, V Jakhetiya, J. F. Qiao, X. Li, W. Lin, and D Thalmann, "Model-based referenceless quality metric of 3d synthesized images using local image description.," *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[16] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov 2006.

[19] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57 – 77, 2015.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] Federica Battisti, Emilie Bosc, Marco Carli, Patrick Le Callet, and Simone Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Processing Image Communication*, vol. 30, no. C, pp. 78–88, 2015.

[22] Yu Zhou, Leida Li, Ke Gu, Yuming Fang, and Weisi Lin, "Quality assessment of 3D synthesized images via disoccluded region discovery," in *IEEE International Conference on Image Processing*, 2016, pp. 1012–1016.